DOCUMENT RESUME

ED 166 713 CS 204 694

AUTHOR Nold, Ellen W.
TITLE The Basics of Research: Evaluation of Writing.
PUB DATE Dec 78
NOTE 12p.; Paper presented at the Annual Meeting of the Modern Language Association of America (93rd, New York City, December 27-30, 1978) ; Best copy available

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Composition (Literary); *Educational Research; *Educational Theories; *English Instruction; Evaluation Methods; Higher Education; Instructional Improvement; Reading Research; Research Design; *Research Methodology; Research Problems; *Student Evaluation

ABSTRACT

Empiricists may be defined as naive fumblers, people who collect and analyze data in the absence of a guiding theory. A theory may be defined as an empirically verifiable explaration of phenomena that predicts behavior. Empirical research may be defined as scholarship that is public and is performed using entities measured in units that can be agreed upon and reproduced. Thus, those who do empirical research must be both well-versed in theory and creative in their designs. Traditional English studies do not provide preparation for doing the theoretical and empirical studies required for charting students' production and comprehension processes or teachers' evaluation processes. The difficulty is that, without an underlying theory, it is not possible to understand the results of research. One result has been that in testing evaluators of writing, important variables have been missed such as the interaction of the reader with the text. Using an interactive theory, at least one researcher has developed some important basic insights into reading and evaluation. (TJ)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Ellen W. Nold

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY.

The Basics of Research: Evaluation of Writing

Talk delivered at the MLA
12-28-78

Ellen W. Nold
Stanford University

BEST COPY AVAILABLE

- ED166713

The title of my talk, listed in the program is: What do evaluators see when given a controlled evaluation task? The answer to my question is "I don't know yet." One year ago when Ed Corbett asked me to speak, I was sure that I would have completed a bit of empirical research on evaluation of writing by now--something along the lines of my previous work. But rather than doing empirical research, I must confess that I have spent the last year thinking. Now thinking and empirical research are not necessarily mutually exclusive, but in my own case, my previous empirical research was not preceded by much in the way of thought besides that given to the mechanics of design. This is the case, I suspect, with most of the empirical studies in our field.

I wish to suggest in this paper not that we avoid empirical research but, paradoxically, that we cease to be empiricists so that we can do our empirical research better. Because this panel addresses empirical studies, I will illustrate my point with past studies and show how more recent ones improve when their researchers stop being empiricists.

In my remarks so far I have used three terms--empiricist, theory, and empirical research--terms that need definition before the full force of my argument can be felt. My Random House Dictionary defines an empiricist as a quack or charlatan. These are strong words to denote someone who collects and analyzes data in the absense of a guiding theory, someone who does so intentionally, knowing that he can mislead himself or others thereby. I know that I did not intend to mislead others in the past; neither do most of my well-meaning empiricist colleagues. So we do not merit the designation quack or charlatan. But we do merit the designation naive fumblers, because naive fumbling is what

C S204696

one does when one is beginning to try to understand a phenomenon. And because empirical research in the production, comprehension, and evaluation of language in all its complexity is a relatively new field, almost all of the empirical research published in the English journals like Research in the Teaching of English is naive fumbling. My own included, but that story is for later.

Let's go on to the meaning of theory. John Warnock, in his landmark article in CCC called "Who's Afraid of Theory", gives several valid definitions. The one he likes is "that abstract conception of the whole by which the relation of part to part and part to whole may be well understood" (Warnock, 1976, p.17). This definition is broad enough to allow taxonomic systems to be called theories--for example James Kinneavy's taxonomy in A Theory of Discourse (1971). But it is too broad for me. I prefer a narrow definition, one that would exclude most, if not all, of the statements that are presently called "theories" by people in English Departments. For the purposes of this paper, a theory is an empirically verifiable explanation of phenomena that predicts behavior. The reliability of certain taxonomies can be empirically verified for instance, by having tutored (or, in the stronger experiment, untutored) subjects place entities into classes. The reliability of their placements can be statistically determined. But taxonomies cannot be theories because taxonomies are statal: they do not predict the development of phenomena or their other behaviors over time. That attention in composition studies is moving from product to process is significant in that the movement signals a ripeness for the development of real theories of composing, comprehending and evaluating texts. And finally, for the sake of this argument, let's define empirical research. Empirical research is different from that scholarship accomplished by most English Department members in that it is public: it is performed using entities measured in units that can be agreed-upon and reproduced. One researcher can test what another researcher has done and compare

his results to that researcher's, providing he has followed that researcher's method and that neither researcher has made errors in specification or measurement. Or taking the same data, another researcher can offer an explanation or analysis that he shows is preferable to the analysis offered by the first researcher because the second analysis explains patterns in the data more consistently or more elegantly. In other fields with long histories of empirical research--like physics, psychology, and economics--it is usually incumbent upon the challenger, however, also to design and successfully perform other research studies that show his analysis to be preferable. Thus, those who do empirical research well must be both well-versed in theory (though not necessarily theoreticians themselves) and creative in their designs, as well as competent to analyze the data that is produced by their research. Because one person is rarely expert in each of these activities, research is often done in teams--a practice becoming more common in our own field.

It is an understatement to say that traditional English studies do not prepare us well to do or even understand the theoretical and empirical studies required for us to make strides in charting the production and comprehension processes of our students, or our own evaluation processes. But understand we must, for the improvement of our instruction vitally depends on these basic insights. It is not enough to let the cognitive psychologists or psycholinguists do all the work for us: they tend to veer off into unprofitable sideroads--they make assumptions that are untenable, impose restrictions that are terrifyingly simplistic.

As we prepare ourselves to do the work required of us by consciously retooling ourselves, by training our graduate students who wish to do research in composition to think in ways different from traditional ways of thinking about English, there are pitfalls we can avoid. One is to produce chronic

naive fumblers: those who may have some competence in design and interpretation--who may know their Latin Squares and error terms (or at least work with others who know them)--but who are content to fumble naively forever in absense of theory. It is not enough to sigh that our field is "preparadigmatic" (in Kuhn's terms) or roughly "pretheoretic" in mine. Because of the stage of the development of our field, empirical researchers must be the prime developers or at least critics of at least working theories if any of their efforts are to be useful. And these theories must be stated, not just implied, in reports of their research.

Let me use as an example of some fairly useless research some of my own. I did it together with Sarah Freedman, whose advances I'll speak about later, but I take full responsibility for its faults. Technically, this research was nifty enough to get nominated for the Braddock award for the most significant piece on composition in 1977 (Nold and Freedman, 1977). Indeed, it was carefully done, and it used a sophisticated multivariate analysis not much seen in articles in Research in the Teaching of English.

In our research Sarah and I tried to determine correlates with quality ratings of essays: what features of the essays correlated with the composite wholistic score given by six readers. Choosing variables to correlate with quality ratings was a seat-of-the-pants activity. We chose some syntactic variables--the first fourteen of the seventeen variables listed on the first page of your handout--and a vocabulary variable, and variables that dummied for number and elaboration of ideas--LONG for much, many, or well-elaborated ideas, SHORT for few or little elaborated ideas. The syntactic variables were chosen because Lester Golub used them to determine his Syntactic Density Score, a measure that was supposed to predict both reading difficulty and writing maturity for elementary school children. The study in itself was a seat-of-the pants exercise, inadequately supported by theory. The vocabulary variable was chosen

because it reflected control of the lexis, and length because it reflected control of the ideas.

Our student wrote a total of 88 essays on four argumentative topics; features of these essays were counted. Six readers were trained and read the 88 essays, giving them quality ratings from 1 to 6, 1 being the highest. In regressing the variables against the quality ratings, we found that five variables together explained 42% of the variance of the quality ratings with a very low probability of error ($p < .001$). These five are marked with a double star on your handout. Only one was positively associated with quality: the percentage of words in free final modifiers. The more words in free final modification a paper had, the more likely it was to be judged of high quality. The other four variables--short essays (those with a number of words one standard deviation below the mean for that essay topic), percentage of common verbs, percentage of verbs with be or have in the modal position or with other modals (should, could, would)--these variables were associated with low quality ratings. About half the variance we could explain was explained by having a short essay --that was very damning. But short essays were also likely to lack free final modification and bristle with verbs with modals. Evidently, argumentation, unsupported assertions on moral grounds (he should have done it) or grounds contrary to fact (he could have done it) are particularly unconvincing.

This naive fumbling occasioned an epiphany for me. Arriving home from the computer center after midnight waving the printout of the final computer run in glee, I asked by husband, who is a statistician and economist (and my patient helpmeet) what he thought of the results. After looking at them, he remarked, "Well, these results are impressive, but you explain only 42% of the variance. Obviously you have left some very important variables out. What does your theory suggest they might be?" "Theory?" I asked. "Yes," he replied, "The only

way these results can properly be interpreted is in relation to a theory they test." "Oh," I remarked.

This revelation sent me back to look at the implied theoretical underpinnings of the most influential empirical studies I had reviewed before undertaking the research reported by Sarah and me in RTE in 1977. Possibly the most important study was done by Diederich, French and Carlton (1961) of ETS fifteen years earlier. To show the amount of progress we made over their study, let me point out that ours explained almost exactly the amount of variance theirs did. Performing factor analysis on the written comments of many raters, they found the comments to cluster into five groups which they named: ideas, mechanics, organization, wording and flavor. They concluded that, since different raters' comments were focused on different factors, variance in the wholistic ratings of one paper by many raters was due primarily to their different weightings of these factors. That is, they assumed that the error arose from the evaluation and not the prior reading of the texts. There occur to me two criticisms of this reasoning, one trivial and one quite important. The first, trivial, criticism is of their assumption that volume of comment indeed reflected the salience of the trait commented upon. Salience is the psychological robustness of the trait commented upon. They never tested this hypothesis experimentally, nor has anyone, to my knowledge. The second, far more important criticism of both the Diederich, French and Carlton study and of mine and Freedman's--and of most other research in evaluation--is the unquestioning assumption of an underlying text-centered theory of reading. In this theory, the reader is the passive recipient of the text: he gets meaning from the text. Everything in the text is available to every reader, who responds to it and (if the reader is also an evaluator) assigns value to it. Since all the evaluators perceive the same text, errors arise not in their perception and comprehension of it but in their evaluation of what they have perceived and comprehended.

For years, research on evaluation has pointed to the inadequacy of this theory of reading, but the empiricists--no theoreticians, they--did not respond to these signals. The unruliness of the teams of evaluators--their horrifying lack of agreement--led practitioners of evaluation to find ways of assuring agreement. In doing so, they controlled for important variables that should have them suspect the theory they unwittingly had adopted: these variables were reader variables like education, experience with students, fatigue variables that had little to do with the weights given factors of a reading.

Perhaps it is some practioners' troubles that have in part led to an increasing realization that an interactive theory of reading is a more powerful one than a text-centered one. In this the interactive view, the reader makes meaning from the text; he has an active role in the process. And because of his different experiences and expectations and immediate situation, he may perceive a text quite differently than another reader.

The variance in evaluation, then, may indeed arise from different weights given various perceived attributes of the writing, but it most probably also arises from differences in the original perceptions and interpretations of the writing. The interactive theory of reading, then, makes our task of understanding evaluation more difficult and complex, since variation can arise both from the evaluation and from the prior reading process.

In addition, accepting the interactive theory of reading as a prior to evaluating means that one cannot legitimately describe only features of texts and assume that these features cause quality ratings (as, for example, Gebhart [1977] does). Rather, an interactive theory of reading suggests that one must also study the readers of the texts to determine, among other things, how their varying knowledge

bases aid them in comprehending texts and how salient various aspects of the text are for them.

Using this interactive theory of reading as a prior to a study of evaluation, Sarah Freedman completed work on her doctoral dissertation in 1977. (It's confusing because our research together actually was done and written up in 1974-5, but was published after a two-year wait in the RTE cue.) Her dissertation research is a landmark because she not only studied her readers--classifying them into Type 1 through 3, depending on their years of education and years of teaching, but, unknown to them, she also systemmatically rewrote the essays they read to test whether they could perceive gross differences in four aspects: content, organization, sentence structure and mechanics. She took actual essays written by students and had them rewritten by research aides using explicit rewrite rules to be either very strong or very weak in these four qualities and presented 12 readers 96 essays on 8 topics that reflected 12 combinations of strengths or weaknesses: for example--strong content, weak organization, weak sentence structure, strong mechanics is one of the 12 possible combinations.

Readers were asked to read all the essays wholistically and then to give a plus or minus for the quality of content, organization, sentence structure and mechanics. The plus or minus scores for each reader were then matched against the rewriter's intentions of strong or weak in each category. Though 10 of 12 reader's perceptions of the rewriters' intentions matched pretty well, two of the readers--both Type 3's--could not divine the rewriters' intentions at all reliably. There appeared to be no pattern to their errors. Because, their coefficient of match was significatnly lower than the other ten, they were dropped from the study.

This finding is significant and merits further study because it means that 2 of 12 teachers of Freshman Composition at a major university--all with average or better ratings as instructors--were unable reliably to perceive even the grossest differences in four distinct aspects of students' texts. But accurate perception of the qualities of texts is an important prerequisite for consistency and fairness in teaching! It also means someday that using some measurable criteria, researchers and testers can predict which people may have trouble reading and hence not be good evaluators.

A second significant finding in this study was related to the salience of the perceived quality of strength or weakness in content, organization, sentence structure or mechanics in the assignment of wholistic grades. Freedman found that the difference between perceived strong and weak content and organization showed up as an average difference in wholistic score of over 1 point in a 4 point range. But the difference in wholistic score between perceived strong and weak sentence structure and mechanics was only two-thirds of a point. An analysis of variance showed that while the perceived content, organization and sentence structure contributed significantly to the wholistic score, perceived mechanics did not.

This short exposition does not do justice to Freedman's study: it contains many other insights that I am at a loss for time to discuss. But I hope you can get a sense of the kinds of insights that can be gained when empirical research is based explicitly on a powerful theory of reading and evaluation.

Of course, Freedman's work is just a first step in the direction of specifying and examining the variables in the complex process of evaluation--much more theoretical work needs to be done before we can hope to identify all

the reader and text variables that we need to examine in order to explain the process of assigning scores--both to ourselves as researchers and to our students--who ultimately bear the weight of our assessments.

Ellen W. Nold
Stanford University

References

Paul Diederich, S. W. French, and S.T. Carleton, Factors in Judgments of Writing Ability. Research bulletin RB-61-15. (Princeton, New Jersey: Educational Testing Service, 1961) ED 002 172.

Sarah W. Freedman, Influences on the Evaluators of Student Writing (Stanford University: Unpublished doctoral dissertation, 1977).

Ann O. Gebhardt, "Writing Quality and Syntax: A Transformational Analysis of Three Prose Samples," Research in the Teaching of English, II, 2, Fall 1977, pp. 211-231.

E.D. Hirsch, Jr., "Measuring the Communicative Effectiveness of Prose," Writing: The Nature, Development and Teaching of Written Communication, eds., C. Frederiksen, M. Whiteman and J. Domenic,(Hillsdale, New Jersey: Lawrence Erlbaum Associates, in press).

Ellen W. Nold and Sarah W. Freedman, "An Analysis of Readers' Responses to Essays," Research in the Teaching of English, II, 2, Fall 1977, pp. 164-174.

John Warnock, "Who's Afraid of Theory?" College Composition and Communication, XXVII, 1, February 1976, pp. 16-20.

The seventeen variables regressed against quality:

Words per T-unit
Subordinate clauses per T-unit
Mean main clause length
Mean subordinate clause length
% prepositions in sample
% possessives in sample
% adverbs in sample
** % finite verbs that have modal auxiliaries
** % finite verbs that have be or have as auxiliaries
% finite verbs in passive voice
% finite verbs in progressive mood
% gerunds, participles, absolutes
** % words in free final modifiers
% words in free medial modifiers
** % finite verbs that are common*
% essays assigned LONG
** % essays assigned SHORT

*be, break, do, feel, follow, find, get, give, go, have, hold, keep, know, look, live, make, mean, need, put, say, see, seem, stop, take, talk, tell, turn, try, want, work, write, use.